

Dealing with Big Data: the Case of Twitter

Erik Tjong Kim Sang (Netherlands eScience Center, Amsterdam)
Antal van den Bosch (Radboud University, Nijmegen)



by SURF & NWO

Netherlands eScience Center

We support and reinforce multidisciplinary and data-intensive research through creative and innovative use of information and communication technology

We distribute project calls which offer both money and expertise

Currently we are involved in 21 projects dealing with various science areas



Netherlands eScience center

The project TwiNL

Twitter: a social network on which participants communicate by exchanging messages of up to 140 characters

With about 8 million messages in Dutch per day covering different topics, this is an interesting data source for researchers

Project task: collect Dutch tweets, offer a search facility for the data and provide different views on search results

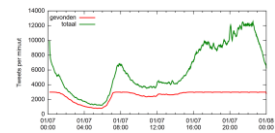
Netherlands eScience center

Collecting the data

We continuously search for common Dutch words on Twitter (keyword tracking)

Additionally we collect all the tweets of most frequent 5000 users in Dutch of the previous month (user tracking)

This generates close to 4 million tweets per day



Netherlands eScience center

Searching the data

A web interface has been built for searching the data: <http://twigs.nl>

Netherlands eScience center

Challenges and solutions

We collected two years of Dutch tweets: about 2 billion (2,000,000,000) tweets with metadata taking up about 5 terabytes (5,000,000,000,000) of disk space

The collection is updated continuously

We want to be able to search both tweet text and the associated metadata

We divided the data in files containing one hour of tweets

Parallel processing on a 530-node Hadoop cluster accessible via Google's MapReduce framework, is used for limiting the search times

Results of earlier searches are kept in a cache

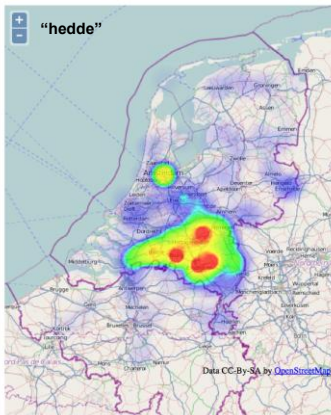
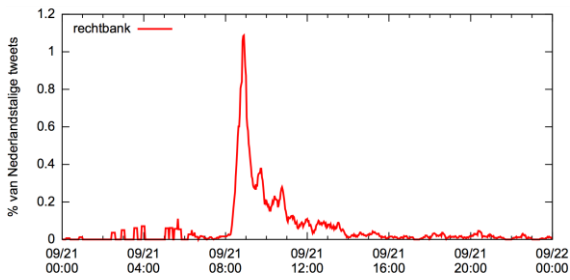
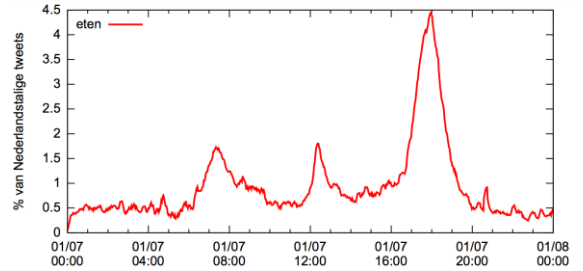
Netherlands eScience center

Views on search results

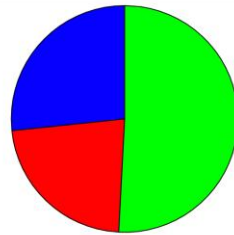
We offer five different views on search results:

- Tweets (restricted)
- Keyword frequency graph
- Keyword usage map
- Vocabulary overview, including sentiment analysis
- User analysis

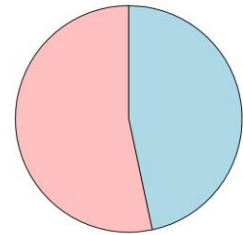
All numbers used for creating the views are available for download



All tweets of Tuesday 4 December 2012

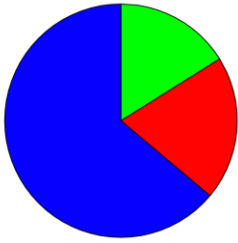


Leeftijd
 < 18 jaar: 50.7%
 18-25 jaar: 22.4%
 > 25 jaar: 26.8%

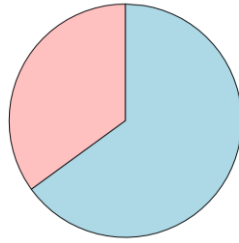


Geslacht
 man: 46.6%
 vrouw: 53.3%

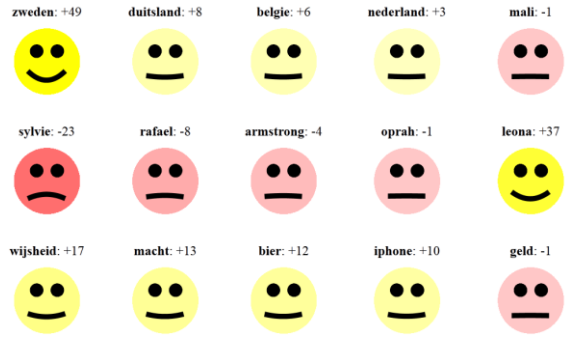
“bultrug”



Leeftijd
 < 18 jaar: 16%
 18-25 jaar: 20%
 > 25 jaar: 64%



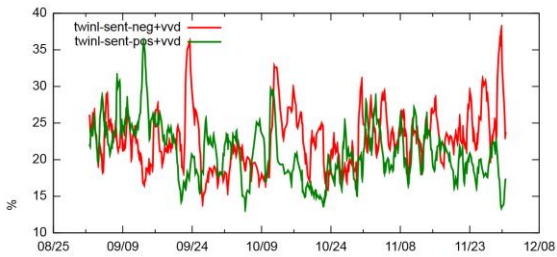
Geslacht
 man: 65%
 vrouw: 35%



Science center



Sentiment for “VVD” in september-november 2012



Science center



Concluding remarks

We have build a web interface for searching Dutch tweets: <http://twiqs.nl>

We use a parallel architecture (Hadoop) rather than an index for storing the underlying data

The interface can be used by university staff and students for exploring Twitter conversations, for example for topic discussions, dialect usage and product sentiment

Case studies relying on this data resource will be presented after this talk

Science center

